AD_____

GRANT NO: DAMD17-94-J-4076

TITLE: Development of a Common Database for Digital Mammography Research

PRINCIPAL INVESTIGATOR(S): Robert M. Nishikawa, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: October 1995

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1995	Annual 15 Sep 94 - 14 Sep 95

4. TITLE AND SUBTITLE
Development of a Common Database for Digital Mammography Research

5. FUNDING NUMBERS
DAMD17-94-J-4076

6. AUTHOR(S)
Robert M. Nishikawa, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Chicago
Chicago, Illinois 60637

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this infrastructure project is to develop a large database of digitized mammograms that will be distributed free of charge to researchers working in all aspects of digital mammography. This database will facilitate and promote rapid development in digital mammography research. The database will consist of 1000 cases subdivided into 5 categories, 4 containing different breast lesions -- masses, microcalcifications, architectural distortions, asymmetric densities (both benign and malignant) -- and one containing normal mammograms. The mammograms will be collected and digitized (0.05-mm pixel size) at two sites: the Universities of Chicago and North Carolina. The database will be stored at the two sites and will be available over internet, and by mail on CD, tape and magneto-optical disks. To date 178 cases have be digitized. Each case consists of index and previous exams (each having four standard views) and up to two special-view mammograms (e.g., magnification views). Another 300 cases have been identified and will be added to the database in the next year. The computer systems for the database have been assembled and are connected to the network. The first release of database should be ready by the end of 1995.

14. SUBJECT TERMS
digital mammography, database, information systems, image analysis, compute-aided diagnosis, image processing breast cancer

15. NUMBER OF PAGES
9

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

*U.S.GPO:1993-0-358-779

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

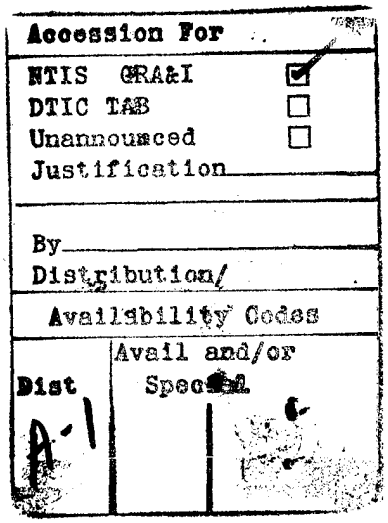
____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



PI - Signature 10/6/95 Date

4. TABLE OF CONTENTS

5. INTRODUCTION

This research is to develop a large database of digitized mammograms that will be distributed free of charge to interested researchers. It is being funded by the USAMRMC as an infrastructure award and as such there it does not represent a research project per se. That is, there is no hypothesis that we are trying to prove. Therefore, this report is structured slightly different than a normal scientific research report -- heavy on the method and light on actual results. In this project, the procedure is the most important component, which is applied continuously in a straightforward manner to achieve the goal of creating the database of mammograms.

5.1 Nature of the Problem

In 1992, the National Cancer Institute identified digital mammography as an important area of research for reducing breast cancer mortality.[1] As a result, there has been a sharp increase in the number of researchers developing computerized methods for analyzing mammograms. This is due in part to the substantial potential benefit from developing an automated computerized system for assisting radiologists in interpreting mammograms. With a large number of investigators developing computerized analysis techniques, the likelihood of an accurate method being developed is high. Unfortunately, a major obstacle to rapid progress in developing a technique is that each investigator uses his or her own set of mammograms (database) to develop and evaluate the performance of his or her technique. As a result, it is not possible to compare the accuracy of different methods because the measured performance is dependent on the cases used for testing.[2] For example, by using "easy" cases for testing, a computer technique would apparently have a higher accuracy than if "hard" cases were used. A common database of mammograms that could be used by all investigators in the field would solve this problem.

5.2. Background: Previous work in the field

At a Biomedical Image Processing meeting held February 1993, in San Jose CA, 12 panelists discussed the design of a common database for research in mammographic image analysis.[3] Two of the panelists are investigators are on this proposal. Important considerations in developing the database are: (a) the cases selected, (b) the digitizer used, (c) organization of the database, (d) associated information to be included with images, (e) "truth" for each case, (f) format of image files, (g) distribution of the database, and (h) rules on using the database.

There have been several small databases released for general use. However, all have several limitations to due to insufficient spatial resolution, insufficient grey-scale resolution, and/or too small a number of cases. The database that we are developing will have none of these limitations.

5.3. Purpose

The purpose of this proposal is to develop a database of digital mammograms that can be used by researchers who (1) are trying to determine the image quality requirements of detectors for digital mammography; (2) are developing image processing techniques to optimize the displayed digital mammogram; (3) are developing computerized methods for analyzing mammograms; (4) are studying the effects of image compression methods on image quality; (5) are developing methods for remote transmission of mammograms; and (6) are studying the relationship between image quality and diagnostic accuracy. This database also could be used as a resource for teaching radiology residents and for testing the performance levels of mammographers.

The specific aims of this proposal are:

1. Collect and digitize 200 cases in each of 5 different categories, mammograms exhibiting: (i) clustered microcalcifications, (ii) masses, (iii) architectural distortions, (iv) asymmetric densities, and (v) no lesions (i.e. normals).

2. Make these cases available to other researchers either over computer network (Internet) or by sending images on computer tape or CD. The database will be distributed as widely as possible so that comparisons of different computerized analysis techniques can be standardized.

5.4. Method of Approch

Task 1: Collect and digitize mammograms, Months 1-48. (See Figure 1.)

a. Retrieve from film library cases with pathologically-proven lesions (clustered microcalcifications, breast masses, architectural distortion, asymmetric densities), 100 cases of each type and 100 normals (cases without lesions) from each site [University of Chicago (UC) and University of North Carolina (UNC)] for a total of 1000 cases during the entire funding period.
b. At each site, digitize retrieved films and outline the location of the lesion in each abnormal image. The outline will be stored together with the images but in a separate file.
c. Send normal cases and asymmetric density cases that were digitized at UC to UNC; and send cases containing masses, microcalcifications, and architectural distortion that were digitized at UNC to UC.
d. Selectively randomize 200 cases for each lesion type into one of two sets (training and testing), based on lesion subtlety. Similarly, selectively randomize 200 normal cases into two sets based on breast density.
e. Place testing set in off-line storage and training cases in on-line storage.
f. On average 250 cases (2500 image -- see text for details) will be done per year for 4 years for a total of 1000 cases (10,000) images.

Task 2: Establish protocol for transmitting database. Months 1-24

a. Test protocols for different modes of transferring data between the UNC and UC (FTP, 8-mm tape, and CD). A data structure designed for portability will be provided to contain the patient text data; this data structure will be made available along with the data to the requesting sites. Use of ACR/NEMA DICOM protocol will be investigated and incorporated as an optional transfer mechanism.

Task 3: Maintain database and distribute cases Months 12-48.

a. Maintain computer, jukebox, and network connection including bug fixes and installation of vendor software updates.
b. Distribute cases via computer network and by mass storage media (tape or CD) as requested.

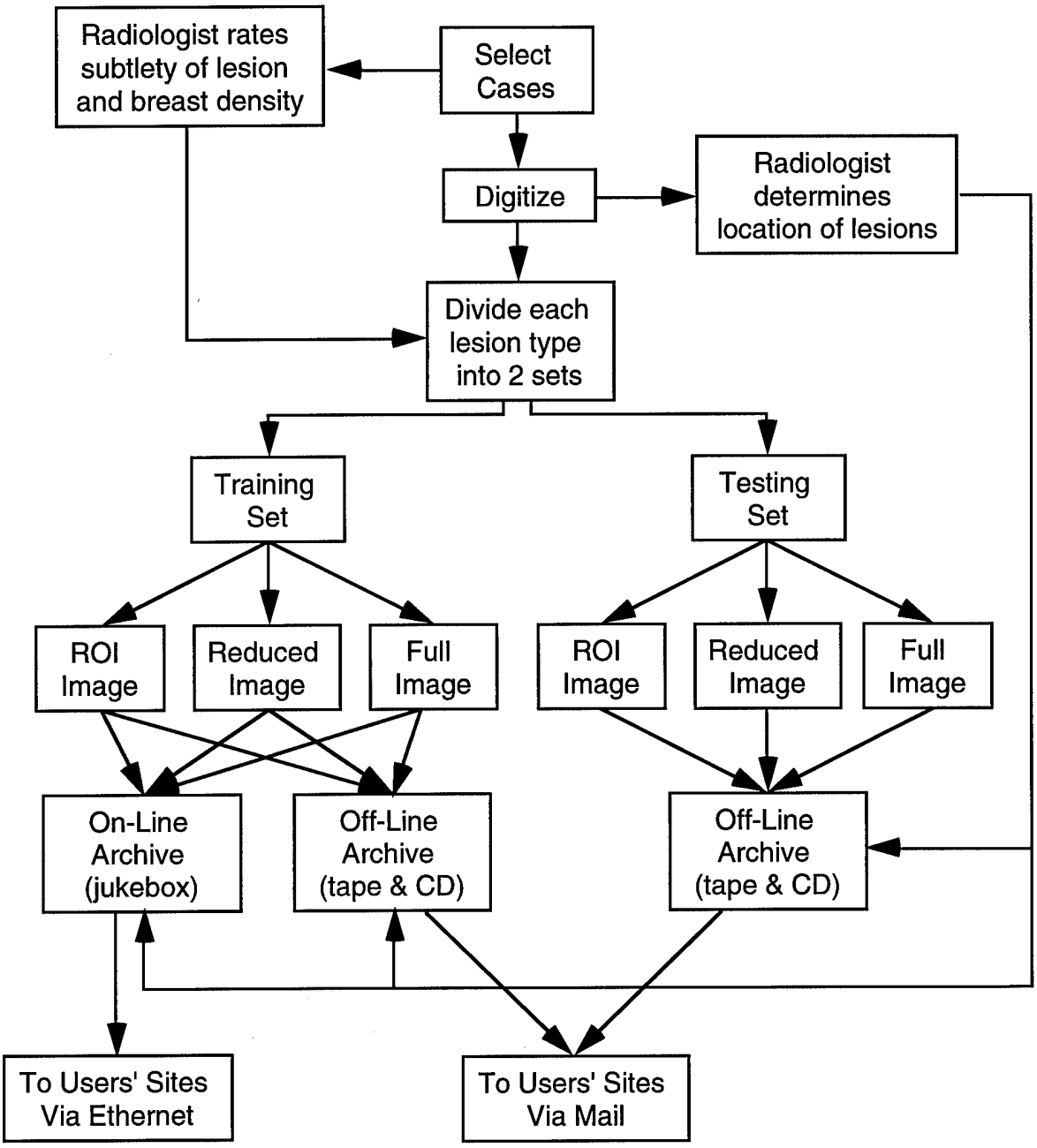


Figure 1. A flowchart of the steps required to collect, digitize, archive, and distribute the mammographic database. The 'Full Image' is the whole digitized mammogram at full resolution. The 'Reduced Image' is a minified version (reduced resolution) of the full image. The 'ROI Image' is a portion of the full image at full resolution.

6. PROGRESS TO DATE

Task 1.

We have retrieved, digitized, classified and filed 178 cases, 111 for the University of Chicago (UofC) and 67 from the University of North Carolina (UNC) as of October 1, 1995. These include lesions from all categories with the majority being masses and microcalcifications. See Table I. The image sublety has been ranked on a 5 point scale (1-5) with 1 being the most difficult to detect. All cases are archived on 8-mm tape.

The computer systems that will hold the database have been purchased and installed. Current capacity of each system is approximately 40 gigabyte (one system at each site). The total capacity of each system will be increased in the third year of the project.

Task 2.

At his point, we have not transferred data between the two sites. This will be done in the second year, when a larger number of cases has been collected. We originally considered the ACR/NEMA (DICOM) image format for our database. However, the ACR/NEMA format does not have a module for mammography, and it would be an extensive project to develop one at this time. Currently, then, we are storing the images as a binary array of numbers with a simple 512-byte header. When an ACR/NEMA mammography module becomes available, it will be easy to convert our files to that format.

Task 3.

Maintenance of the database and distribution of the database are at a minumum currently. These tasks will become important in the next and subsequent years as cases go "on-line".

7. CONCLUSIONS

The develpment of a common database of mammograms for digital mammography research is well underway. The first release of a portion of the database is anticipated for the end of 1995. This release will include 100 cases of clustered microcalcifications.

A database of mammograms would also be useful for investigators doing research in other areas of digital mammography, such as x-ray detector development, telemammography, image compression, and image processing. For example, questions such as the required spatial resolution of a digital mammogram can be answered in part by conducting observer studies using the mammograms from the database displayed at different resolutions. Furthermore, the database would provide an excellent source of cases that could be used for teaching purposes.

8. REFERENCES

1. F. Shtern, "Digital mammography and related technologies: A perspective from the National Cancer Institute," Radiology 183, 629-630 (1992).

2. R. M. Nishikawa, M. L. Giger, K. Doi, F.-F. Yin, C. J. Vyborny and R. A. Schmidt, "Effect of case selection on the performance of computer-aided detection schemes," Medical Physics 21, 265-269, (1994).

3. F. Shtern, "Panel discussion: Design of a common database for research in mammogram image analysis," Proc. SPIE 1905, 534-551 (1993).

Table I. Breakdown of cases in the database as of October 1/95.

Type of Lesion	Pathology	Subtlety	# of Cases
Mass	Malignant	1	17
Mass	Benign	1	2
Mass	Malignant	2	12
Mass	Benign	2	8
Mass	Malignant	3	11
Mass	Benign	3	12
Microcalcifications	Malignant	1	14
Microcalcifications	Benign	1	13
Microcalcifications	Malignant	2	18
Microcalcifications	Benign	2	12
Microcalcifications	Malignant	3	16
Microcalcifications	Benign	3	13
Asymmetric Density	Malignant	1	7
Asymmetric Density	Benign	1	0
Asymmetric Density	Malignant	2	6
Asymmetric Density	Benign	2	1
Asymmetric Density	Malignant	3	3
Asymmetric Density	Benign	3	1
Architectural Distortion	Malignant	1	5
Architectural Distortion	Benign	1	1
Architectural Distortion	Malignant	2	3
Architectural Distortion	Benign	2	1
Architectural Distortion	Malignant	3	1
Architectural Distortion	Benign	3	0
Normal	-	-	1
TOTAL			143